7.02.07 - Outras Sociologias Específicas.

APRENDIZADO DE MÁQUINA COMO MEDIAÇÃO TÉCNICA COMPUTACIONAL: VIÉS DE GÊNERO NO PROCESSAMENTO AUTOMÁTICO DE TEXTOS RELACIONANDO PRONOMES E PROFISSÕES PELO ALGORITMO BERT.

Rafael Gonçalves[1], Pedro P. Ferreira[2]
1. Estudante da Faculdade de Engenharia Elétrica e de Computação da Unicamp.
2. Professor do IFCH-Unicamp - Departamento de Sociologia/Orientador.

Resumo

Frequentemente, as tecnologias são consideradas ferramentas utilitárias que não participam da produção de valores na sociedade; mas, paralelamente, é comum compará-las aos humanos, exagerando suas capacidades. Em relação às tecnologias de aprendizado de máquina, essa representação contraditória se intensifica, considerando-as formas neutras e objetivas de processar dados e, por isso, alternativas mais capazes para a tomada de decisão do que agentes humanos. Alternativamente, este trabalho considera tecnologia e humanidade como campos complementares e faz uma análise sociotécnica do algoritmo de aprendizado de máquina BERT, desenvolvido e utilizado pela Google. Especificamente, objetivou-se investigar a produção de viés de gênero no processamento de frases relacionando pronomes e profissões pelo BERT. O trabalho conclui que o BERT atua favorecendo uma visão sexista e estereotipada da relação gênero-profissão e que, por isso, não deve ser considerado uma ferramenta neutra e objetiva.

Palavras-chave: Sociologia da tecnologia; inteligência artificial; Google.

Apoio financeiro: CNPq.

Trabalho selecionado para a JNIC: Unicamp.

Introdução

As tecnologias estão cada vez mais inseridas nas interações sociais, por isso entender o papel que elas ocupam na sociedade é uma questão fundamental da contemporaneidade. Para Simondon (2020a), esse debate está marcado por duas atitudes contraditórias em relação aos objetos técnicos: eles são considerados ferramentas neutras que não participariam da reprodução de valores na sociedade; mas, ao mesmo tempo, são tidos como concorrentes do humano.

Essa contradição se intensifica no campo da inteligência artificial (IA): por um lado, os algoritmos de aprendizado de máquina são apresentados como técnicas utilitárias de processamento de dados “brutos”; por outro, suas capacidades são frequentemente extrapoladas ao comparar a performance destas com a dos humanos, por vezes considerando-as como possíveis sucessoras deste. Elish e boyd (2018) apresentam que, desde seu surgimento na década de 1950, o termo “inteligência artificial” esteve ligado com ideais de capacidades sobre-humanas, os quais persistem hoje. A mídia tem um importante papel na legitimação desse discurso (SCHWARTZ, 2018), mas isso também ocorre de forma mais sutil dentro do próprio campo de pesquisa, por exemplo ao se considerar que “em muitos casos, IA pode reduzir a interpretação subjetiva de dados por humanos (…)” (SILBERG; MANYIKA, 2019, p. 2).

De forma diferente, é possível considerar que humanidade e tecnologia são complementares (SIMONDON, 2020a) e que IA deve ser entendido como um conceito sociotécnico (ELISH; BOYD, 2018). A partir disso, propomos analisar o aprendizado de máquina como *mediação técnica* (LATOUR, 1994), processo através do qual ações de outros momentos, de outros espaços e de outros agentes continuam a agir aqui e agora, na condição de serem alteradas, traduzidas, delegadas ou deslocadas por e para agentes não-humanos (LATOUR, 1994, p. 50). Assim, o objetivo deste trabalho é explorar a ação social da tecnologia BERT – desenvolvida e utilizada pela empresa Google – por meio da análise da produção de viés de gênero no processamento automático de textos. Dessa forma, pretende-se apresentar um caso particular de agência algorítmica para argumentar que as tecnologias de aprendizado de máquina não devem ser consideradas formas neutras e objetivas de processar informação.

Metodologia

A agência dos algoritmos de aprendizado de máquina é muitas vezes percebida no que se convencionou chamar de “viés”, ou seja, tendências não intencionais nos resultados e que, em muitos casos, são entendidas como problemáticas do ponto de vista ético ou político. Baseado em um trabalho similar que analisou viés de gênero na tradução automática do Google Search (PRATES; AVELAR; LAMB, 2019), este trabalho buscou identificar viés de gênero no algoritmo BERT por meio da análise das relações entre pronomes e profissões geradas pelo algoritmo.

O BERT (DEVLIN et al., 2019) é um algoritmo (ou modelo) de aprendizado de máquina desenvolvido pela Google que visa criar representações de um dado textual de entrada que podem, posteriormente, ser usadas na resolução de diversas tarefas de processamento de linguagem natural (PLN). Além de ser um algoritmo muito conhecido da área de PLN, recentemente, a empresa Google anunciou que o BERT seria usado em quase todas as buscas em inglês que fossem inseridas em seu principal produto, o Google Search

(GOOGLE, 2020). Segundo a empresa, o uso do algoritmo "seria particularmente útil para entender a intenção por trás da pergunta" e a incorporação do BERT teria "ajudado a melhorar as buscas em uma escala massiva, impactando 1 em cada 10 buscas em inglês nos EUA" (GOOGLE, 2020).

Uma das tarefas que o BERT é capaz de resolver é a modelagem de linguagem mascarada, em que o algoritmo é treinado para inferir palavras faltantes em uma frase de entrada. Nessa configuração, o BERT recebe como entrada uma lista de *tokens* (palavras e acentuação) que representam uma frase. Estes são separados entre *alvo* (palavras mascaradas que devem ser inferidas) e *contexto* (outros *tokens* que servirão de informação para o algoritmo). A partir do processamento da entrada, o algoritmo gera como saída uma lista de palavras-candidato ranqueadas por probabilidades inferidas pelo próprio modelo.

Tendo em vista esse modo de funcionamento do algoritmo BERT, elaborou-se uma lista de dez palavras em inglês representando profissões ou conjuntos de profissões (*biologist, ceo, doctor, electrical engineer, electrician, engineer, lawyer, nurse, police officer* e *sociologist*) e propôs-se a estrutura "MÁSCARA is a/an PROFISSÃO" a fim de analisar os resultados para os pronomes *he* e *she* (exemplo na tabela 1). Os experimentos foram realizados utilizando um simulador online do algoritmo BERT (BIU, [s.d.])

Entrada: <?> is an engineer.

Saída:
1: he – 81.380939%
2: she – 9.814042%
3: thomas – 0.072868%
4: david – 0.068877%
5: singh – 0.068845%

Tabela 1: Exemplo de experimento realizado. <?> representa a palavra alvo. Ênfase nos resultados com pronomes com marcação de gênero. Mostrando apenas as 5 saídas com maiores probabilidades inferidas.

Resultados e Discussão

A partir dos dados empíricos obtidos (síntese na tabela 2), foi possível perceber uma tendência geral de privilegiar o pronome masculino *he* em relação ao pronome feminino *she*. Esse resultado é similar ao viés constatado em (PRATES; AVELAR; LAMB, 2019), segundo o qual o Google Translate privilegiaria traduções de frases sem marcação de gênero em uma língua para frases no masculino para outra língua.

Profissão	**P(he) [%]**	**P(she) [%]**
Biologist	60,3	28,5
CEO	61,9	15,3
Doctor	61,7	21,2
Electrical engineer	81,2	10,4
Electrician	67,0	4,5
Engineer	81,4	9,8
Lawyer	75,9	18,2
Nurse	**3,0**	**69,8**
Police officer	68,1	16,9
Sociologist	69,1	24,1

Tabela 2: : Síntese dos resultados obtidos. Para cada profissão, P(he) e P(she) *são as probabilidades inferidas pelo modelo para os pronomes* he *e* she *respectivamente. Ênfase na profissão enfermeira/o (*nurse*), que foi a única cuja probabilidade para o pronome feminino* she *foi maior que para o pronome masculino* he*.*

Das profissões avaliadas, a única exceção para essa tendência de privilegiar o masculino foi em relação à profissão enfermagem (*nurse*), o que indica a reprodução de uma relação bastante estereotipada entre gênero e profissão. Dessa forma, foi possível constatar a produção de um viés que privilegiaria o masculino e a estereotipia no processamento automático de texto pelo BERT.

Da perspectiva de Simondon (2020a), o automatismo é um grau baixo de perfeição técnica, pois automatizar uma máquina diminui suas possibilidades de funcionamento em relação a uma máquina aberta, em interação com o humano (os funcionamentos são predeterminados). Dessa forma, a adaptação automática de uma máquina estaria necessariamente atrelada com "condutas estereotipadas" e com "uma ligação estreita com um meio determinado" (SIMONDON, 2020b, p. 529). Nesse mesmo sentido, Pasquinelli e Joler (2021)

apontam que a produção de vieses é uma característica constitutiva das técnicas de aprendizado de máquina e que, como formas contemporâneas de automatismo, elas promoveriam a discriminação social e a perda da diversidade cultural.

De acordo com Pasquinelli e Joler (2021, p. 1266), os "dados são a fonte inicial de valor e inteligência". Isso ocorre, pois os algoritmos de aprendizado de máquina funcionam em dois momentos diferentes: *treinamento* e *inferência*. Para que o BERT seja capaz de inferir palavras faltantes em frases nunca antes vistas (inferência), ele deve ser configurado previamente (treinamento) – diagramas na figura 1. No treinamento, frases completas retiradas de uma base de dados são utilizadas como entrada do modelo. Em cada frase, uma palavra é escolhida como alvo e salva como rótulo. Então, na frase de entrada, a palavra alvo é substituída por um *token* de máscara (80% das vezes), deixada inalterada (10% das vezes) ou substituída por um *token* aleatório (10% das vezes). Por exemplo: a frase "my dog is hairy" com *token* alvo *hairy* seria substituída por "my dog is MÁSCARA", "my dog is hairy" e "my dog is apple", respectivamente (DEVLIN et al., 2019). O objetivo do algoritmo no treinamento é inferir corretamente o *token* alvo (no caso do exemplo: *hairy*). Inicialmente, o modelo produzirá resultados incorretos. Esses resultados incorretos são comparados com o rótulo e a diferença (erro) é utilizada para configurar o modelo automaticamente. Após o processamento de grandes quantidades de frases, o modelo passa a inferir corretamente o alvo (diz-se que ele "aprendeu").

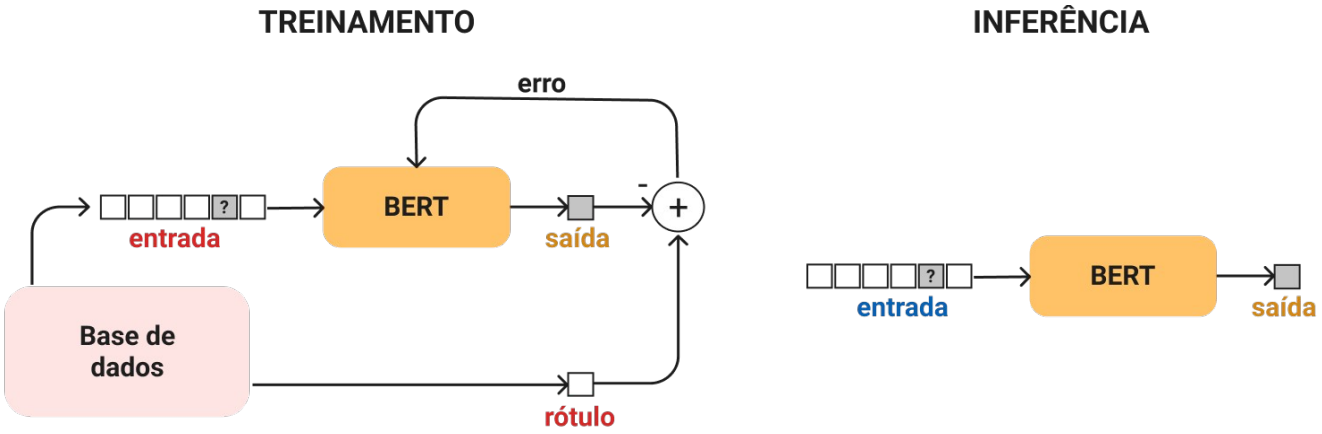

Figura 1: Momentos de funcionamento do algoritmo BERT. À esquerda, diagrama do fluxo de treinamento. À direita, diagrama do fluxo de inferência.

No BERT, os *corpora* utilizados como base de dados de treinamento são o BooksCorpus (800 milhões de palavras) e a Wikipedia em inglês (2500 milhões de palavras) (DEVLIN et al., 2019). O primeiro é composto por livros disponibilizados na internet – com gêneros diversos, incluindo "fantasia" e "ficção científica" (ZHU et al., 2015) – e o segundo contém as páginas em inglês da Wikipedia, com exceção de listas, tabelas e cabeçalhos (DEVLIN et al., 2019).

É evidente que textos literários incorporam valores de seus escritores e escritoras e não podem ser considerados neutros, mas, principalmente, textos da Wikipedia também não devem ser considerados isentos de valores. Koerner (2020) mostra que, apesar da proposta de democratização do conhecimento, existem vieses na produção do conteúdo da Wikipedia, de forma que há uma predominância de conteúdos alinhados com perspectivas hegemônicas. Em especial, sobre viés de gênero: "pessoas e conhecimentos publicados em materiais escritos são largamente de uma perspectiva branca e masculina [*are largely white and male*]. (…) Existe muito menos conteúdo sobre mulheres na Wikipedia do que existe sobre homens (...)" (KOERNER, 2020, p. 317).

Esse desbalanço é incorporado pelo BERT através do que Pasquinelli e Joler (2021) chamam de *viés histórico* – a incorporação de vieses já presentes na sociedade – e *viés de base de dados* – a criação de novos vieses durante a criação da base de dados. Além disso, o próprio processo de compressão de informação no treinamento – a incorporação do conhecimento de toda base de dados na estrutura do modelo – promove um *viés algorítmico,* a amplificação dos vieses histórico e de base de dados pelo algoritmo de aprendizado de máquina (PASQUINELLI; JOLER, 2021).

Conclusões

O discurso atual sobre inteligência artificial, atribui a ela uma posição de neutralidade ao considerar que tecnologias de aprendizado de máquina seriam formas objetivas de processar dados. Ancorados em uma literatura que reconhece que tecnologia e humanidade não são campos totalmente distintos e separáveis, foram realizados experimentos a fim de verificar a ação social do algoritmo BERT em uma tarefa que permite avaliar correlações entre gênero e profissão nos resultados do modelo.

Foi constatado uma tendência geral no algoritmo de privilegiar o pronome masculino. Além disso, tendo em vista que a única exceção que produziu probabilidades maiores para o pronome feminino foi para a profissão enfermagem, os experimentos indicaram a existência de uma representação estereotipada da relação gênero-profissão no modelo. Isso pode ser explicado pelo fato de que o conhecimento gerado pelo algoritmo

provém do processamento de exemplos em uma base de dados. No caso do BERT, a base de dados é composta por exemplos de dois *corpora*: o BooksCorpus, coleção de livros disponíveis na internet, e a Wikipedia em inglês. O fato de que textos literários incorporam valores de seus autores e autoras não é surpreendente, mas a Wikipedia também não pode ser considerada uma fonte neutra de informação, pois seus conteúdos estão marcados, entre outras, por tendências em relação a gênero. Essas tendências presentes nos dados são incorporadas e amplificadas no modelo (vieses histórico, de base de dados e algorítmico). Assim sendo, a utilização do BERT não é uma forma neutra e objetiva de processar informação, mas um processo de mediação, em que o objetivo de inferir uma palavra faltante é negociado com os objetivos e tendências de outros elementos: as bases de dados, escritores de conteúdo da Wikipedia, a própria estrutura do algoritmo, entre outros.

Dessa forma, aprendizado de máquina deve ser entendido como um processo de mediação técnica computacional, pois a utilização destes algoritmos implica a mobilização de objetivos, tendências e vieses de outros agentes – em outros momentos e em outras localidades – seja na constituição do algoritmo, na criação da base de dados ou na sociedade em que ele está inserido. Em outras palavras, algoritmos de aprendizado de máquina não devem ser considerados ferramentas neutras e objetivas, pois sua utilização implica uma agência técnica computacional. Em especial, no caso do BERT, as tendências constatadas – e em conjunto com um discurso que atribui neutralidade e objetividade a tais tecnologias – age socialmente de forma a legitimar o sexismo e o machismo. Assim, constatamos que é de suma importância uma análise sociotécnica que considere a agência dos algoritmos e dos dados, seja nos processos de uso, seja nos processos de construção das tecnologias de aprendizado de máquina.

Referências bibliográficas

BIU. **BERT Language Model Demo**. Disponível em: <https://nlp.biu.ac.il/~ohadr/bert/>. Acesso em: 19 ago. 2021.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv:1810.04805 [cs]**, 24 maio 2019.

ELISH, M. C.; BOYD, DANAH. Situating methods in the magic of Big Data and AI. **Communication Monographs**, v. 85, n. 1, p. 57–80, 2 jan. 2018.

GOOGLE. **Search On 2020**. Disponível em: <https://searchon.withgoogle.com/>. Acesso em: 21 ago. 2021.

KOERNER, J. Wikipedia Has a Bias Problem. In: REAGLE, J.; KOERNER, J. (Eds.). . **Wikipedia@ 20: Stories of an Incomplete Revolution**. [s.l.] The MIT Press, 2020. p. 311–321.

LATOUR, B. On technical mediation. **Common Knowledge**, v. 3, n. 2, p. 29–64, 1994.

PASQUINELLI, M.; JOLER, V. The Nooscope manifested: AI as instrument of knowledge extractivism. **AI & SOCIETY**, v. 36, n. 4, p. 1263–1280, 1 dez. 2021.

PRATES, M. O. R.; AVELAR, P. H.; LAMB, L. C. Assessing gender bias in machine translation: a case study with Google Translate. **Neural Computing and Applications**, 27 mar. 2019.

SCHWARTZ, O. **"The discourse is unhinged": how the media gets AI alarmingly wrong**. Disponível em: <http://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong>. Acesso em: 26 ago. 2021.

SILBERG, J.; MANYIKA, J. Notes from the AI frontier: Tackling bias in AI (and in humans). **McKinsey Global Institute (June 2019)**, 2019.

SIMONDON, G. **Do modo de existência dos objetos técnicos**. Tradução: Vera Ribeiro. 1. ed. Rio de Janeiro: Editora Contraponto, 2020a.

SIMONDON, G. Nota complementar sobre as consequências da noção de individuação. In: **A individuação à luz das noções de forma e de informação**. Tradução: Luís Eduardo Ponciano Aragon; Tradução: Guilherme Ivo. 1. ed. São Paulo: Editora 34, 2020b. p. 507–545.

ZHU, Y. et al. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. **arXiv:1506.06724 [cs]**, 22 jun. 2015.